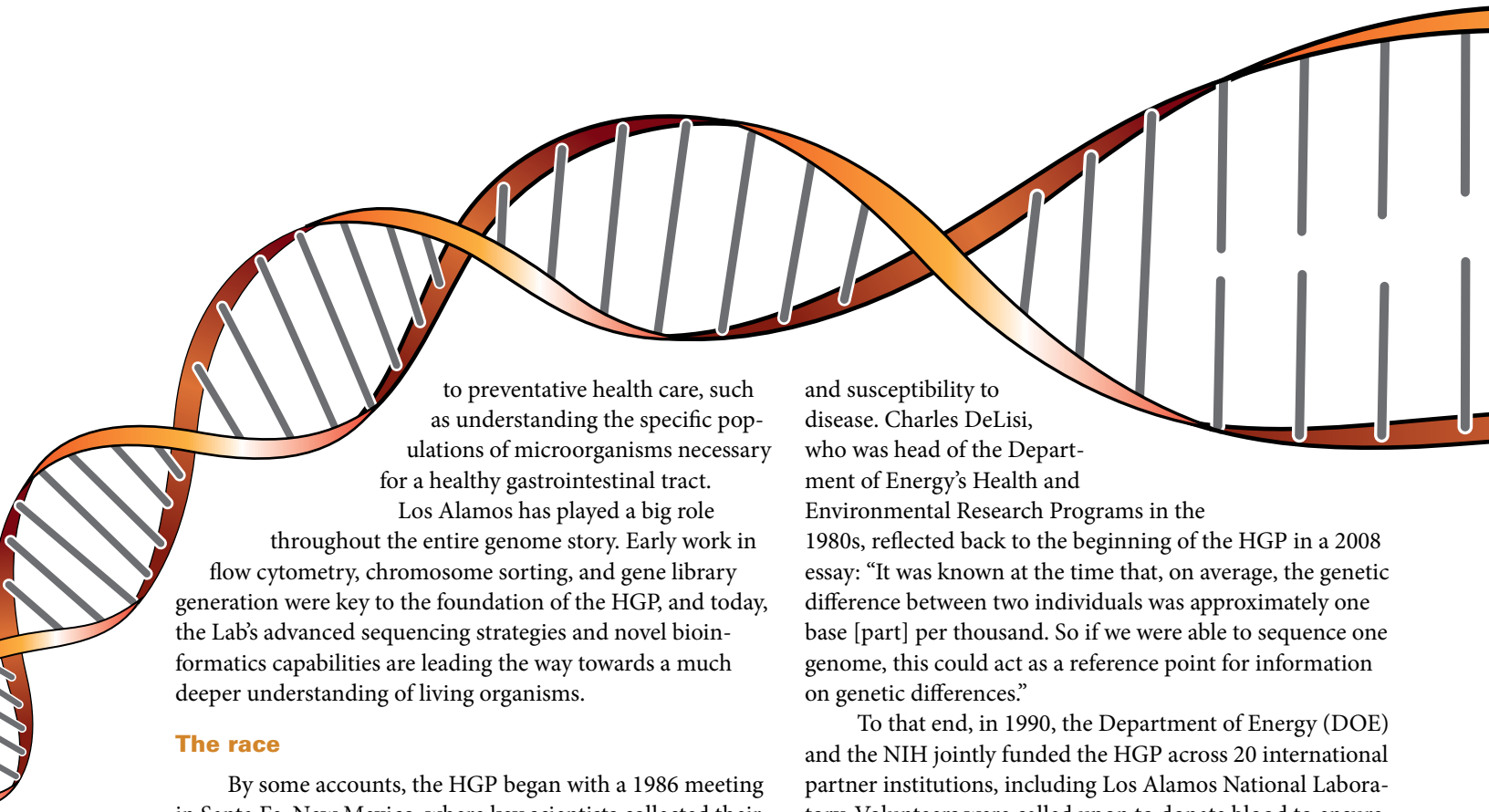
# Unraveling life four letters at a time

*Ten years ago marked the official end of the Human Genome Project, but really it was just the beginning…*

On April 24, 2003, the sequence of the human genetic code, or genome, was published, signifying the conclusion of the Human Genome Project (HGP). A major international effort, the project cost nearly $3 billion and was expected to take 15 years but finished two years early. It is a fascinating story of achievement and even a little drama.

The biggest achievement, however, may not reside in the actual map of the human genetic code, but in the genomics revolution that followed. In the 10 years since the map's completion, genetic sequencing has advanced at a phenomenal pace, revolutionizing the reach of biological research. The technology has become exponentially faster and orders of magnitude less expensive, and advances in bioinformatics are making it possible to thoroughly understand and analyze the mountains of genetic data now available. On the horizon are endless possibilities—from personalized medicine, such as cancer treatment tailored to the distinctive genetic fingerprint of each patient's tumor,



to preventative health care, such as understanding the specific populations of microorganisms necessary for a healthy gastrointestinal tract.

Los Alamos has played a big role throughout the entire genome story. Early work in flow cytometry, chromosome sorting, and gene library generation were key to the foundation of the HGP, and today, the Lab's advanced sequencing strategies and novel bioinformatics capabilities are leading the way towards a much deeper understanding of living organisms.

## The race

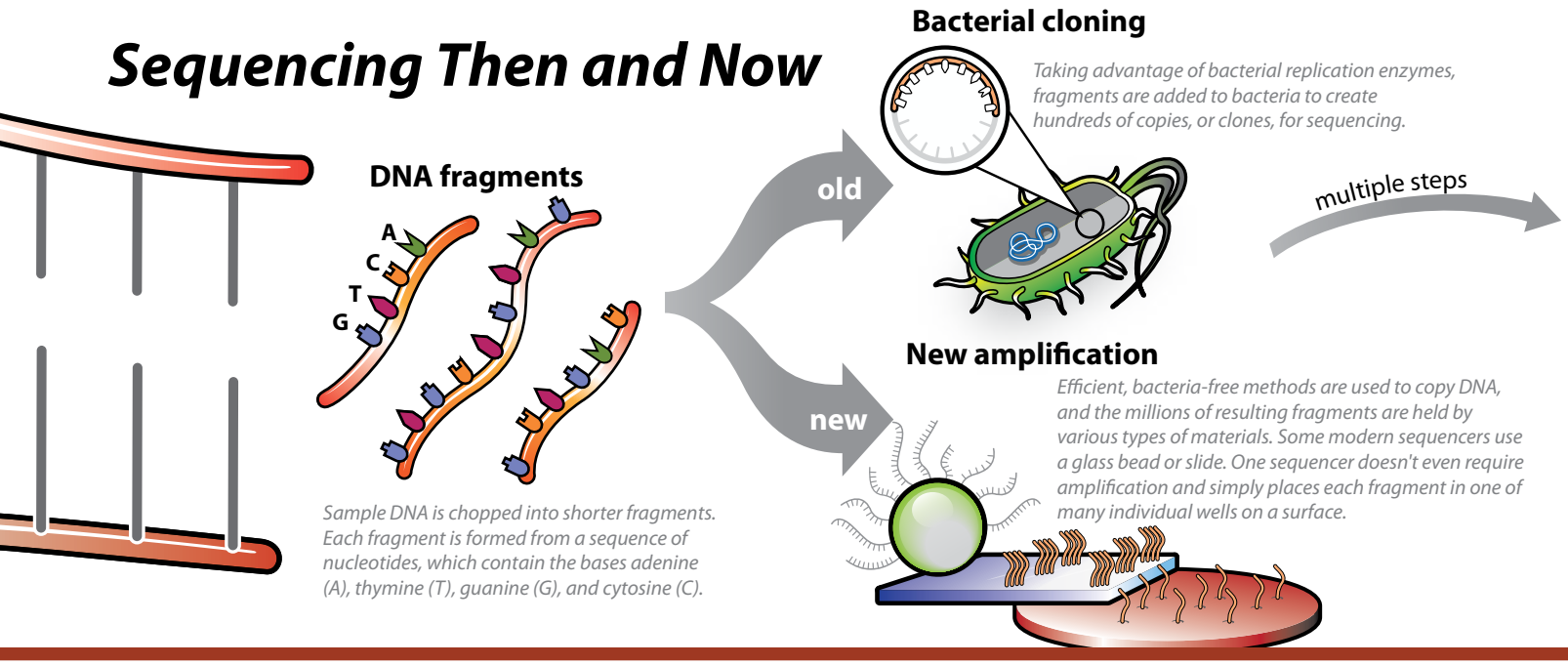
By some accounts, the HGP began with a 1986 meeting in Santa Fe, New Mexico, where key scientists collected their thoughts about why the government should fund the monumental endeavor to map and sequence all the DNA it takes to make a human. Scientists in Los Alamos had a long history of experience in this area, beginning with studying the mutagenetic effects of radiation on cells and leading to the National Laboratory Gene Library Project, in which they isolated, cloned, and packaged chromosomes into libraries for use by researchers worldwide. They accomplished this using flow cytometers, which were invented at the Laboratory [*see Los Alamos Firsts on the inside front cover of this issue of* 1663]. Also foundational to the HGP was Los Alamos's development of GenBank, now managed by the National Institutes of Health (NIH), a public database for genetic sequences.

"Leading up to the HGP, it was a very exciting time here at Los Alamos because we were doing stuff that very few other places in the world were doing," says Los Alamos biologist Jon Longmire.

At that time (the 1980s), a fair amount was already known about human genetics from various experiments and animal studies —multiple gene mutations had been linked to disease and localized to a particular chromosome. However, scientists knew there was much more to be learned about the mechanisms of inheritance, the functions of the human body, and susceptibility to disease. Charles DeLisi, who was head of the Department of Energy's Health and Environmental Research Programs in the 1980s, reflected back to the beginning of the HGP in a 2008 essay: "It was known at the time that, on average, the genetic difference between two individuals was approximately one base [part] per thousand. So if we were able to sequence one genome, this could act as a reference point for information on genetic differences."

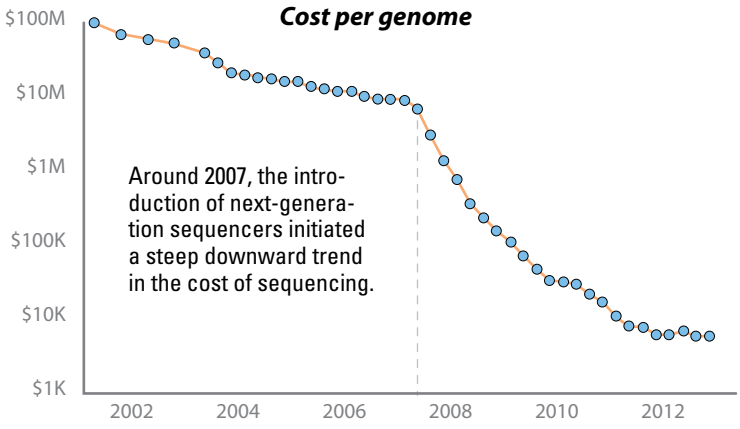
To that end, in 1990, the Department of Energy (DOE) and the NIH jointly funded the HGP across 20 international partner institutions, including Los Alamos National Laboratory. Volunteers were called upon to donate blood to ensure a diverse pool of starting material, the logic being that since human genomes are mostly identical, chromosome samples from a handful of individuals would give researchers a fairly accurate sense of an average person. Los Alamos's role was to create chromosome libraries from the donated material, with which each partner lab would map and sequence its assigned chromosome; Los Alamos would investigate chromosome 16 and part of chromosome 5. The HGP was expected to take 15 years using a laborious, incremental process that involved creating a physical map of each chromosome and sequencing and analyzing small sections of it in their correct order.

In 1998, the effort faced a challenger. A scientist named Craig Venter, who had previously worked with the NIH, created a company called Celera and proposed to complete the task in just three years using an alternative strategy that he had been using for microbial genomes. His approach broke up the entire genome (all chromosomes) into random small pieces, sequenced them, and then relied heavily on new computer algorithms to put the pieces back together in the correct order—a process known as whole-genome shotgun sequencing. Venter's method was far ahead of its time, and few believed it would work for such a large genome. However, it is now the basis of all modern sequencing.




Both public and private efforts published draft genomes in the top two scientific journals during the same week in 2001. The "race," effectively a tie, had been complicated by proprietary concerns, namely that the public effort was openly publishing its data while Celera was not, and questions abounded about patenting genes. The Clinton administration got involved and in March 2000 announced that the genome could not be patented. The "final" high-quality sequence was published in 2003, but in reality it remains a work in progress as many details are still being resolved.

By the end of the project, two breakthroughs had been made. Whole-genome shotgun sequencing had been proven successful, and a high-quality map of the human genome had been made. Initial analysis showed that the genome included about 30,000 genes widely distributed among many repetitive regions. Also found were many clues about recombinations and modifications that contribute to diversity. However, it was clear there would be much more to learn, and scientists are still investigating the complex roles of non-coding regions (sections of DNA that do not code for proteins as sections containing genes do).

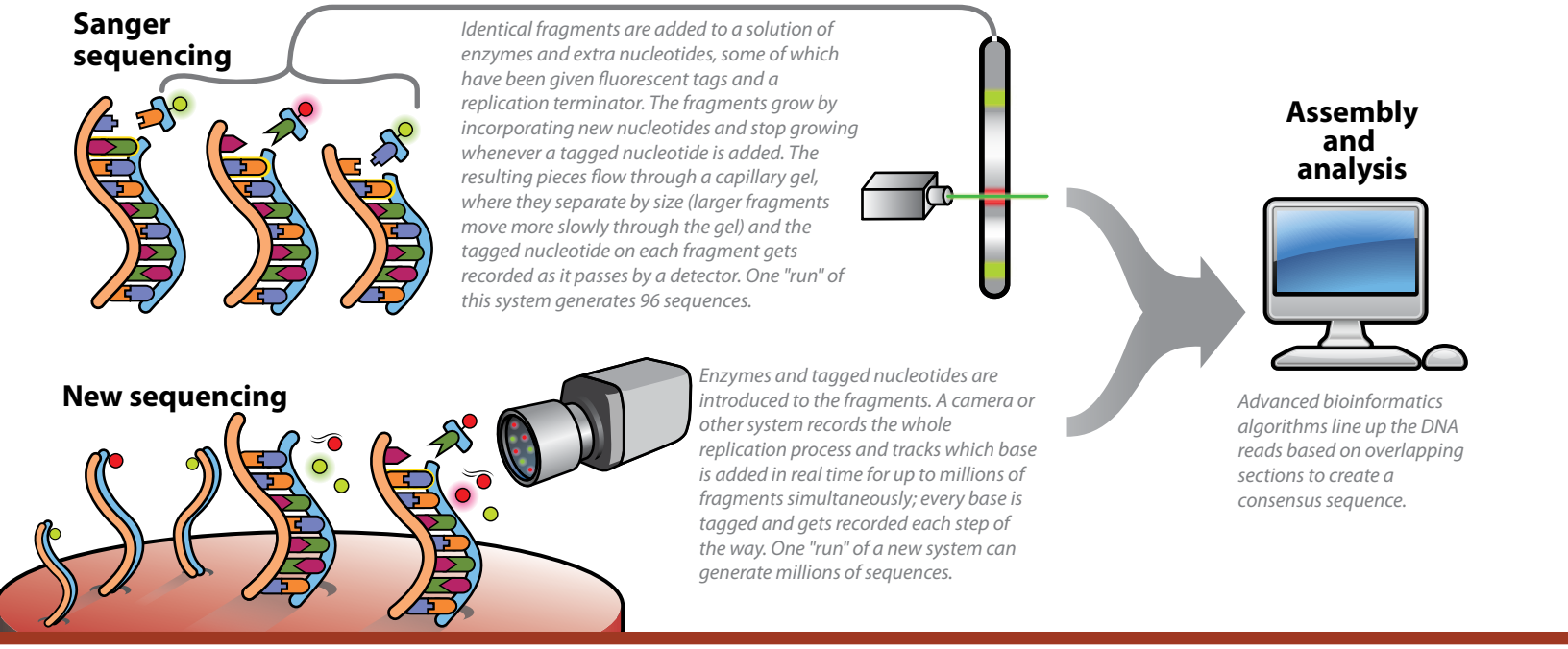



## GATTACA, literally

In the 10 years since the publication of the human genome, major research institutions, such as DOE- and NIH-funded labs, have worked to understand how sequence data translates into an organism's characteristics and functions, while industry revolutionized the technology.

The now familiar "twisted ladder" structure of DNA was discovered in 1953. Its "rails" are made of alternating phosphate and sugar groups, and its "rungs" consist of nitrogenous bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The bases have a very specific binding protocol: A only binds to T, and C only binds to G. With only four bases in this code, the order, or sequence, of the bases is critical to understanding the meaning of the DNA. Methods of determining the sequence of bases began to arrive in the 1970s and the most well-known method, "Sanger sequencing"—developed by Frederick Sanger in 1977—was the primary technology used for the HGP and other projects for more than 20 years.

Most sequencing methods rely on the idea of mimicking DNA replication. During growth, cells multiply in number and must replicate their DNA as part of the process. When DNA is replicated, the ladder splits open into two template strands, and specialized enzymes called polymerases build new strands on the templates using free-floating subunits called nucleotides, each of which is a base plus phosphate and sugar—half a rung and a segment of rail. Laboratory sequencing methods strive to mimic this process and




identify which base is added when the new strand is built; the template strand is then inferred because of the binding protocol. For example, if a T is added, then the template must have an A at that location. However, spying on the activity of polymerases is difficult, so various sequencing strategies have been developed to make the task more tractable.

The first step in sequencing is to isolate and then shear the DNA into smaller, more manageable fragments. Next, most methods involve amplification, or making copies of the fragments. The identical fragments are sequenced and the results integrated to arrive at a reliable consensus sequence.

For traditional Sanger sequencing, the copied DNA fragments are exposed to polymerase enzymes, nucleotides, and a small number of tagged nucleotides that stop the replication process. (Early sequencing methods used radioactive tags, but in the 1980s, fluorescent tags were introduced.) The result is a large collection of pieces of identical DNA of different lengths, where each piece ends with a tagged nucleotide. The pieces then flow through a gel designed to separate them by size (larger fragments move more slowly through the gel), and the fluorescent tags get recorded as they pass by a laser detector. Together, the fragment sizes (organized by the gel) and nucleotide tags spell out the genetic sequence—fourth position T, fifth position C, and so on.

Modern, next-generation sequencing uses various approaches to parallelize the procedure. The result is a much cheaper process. Instead of bacterial cloning, new sequencers use more efficient methods of amplification, and some don't even need to amplify fragments at all. New sequencers use multiple strategies for tagging and detecting nucleotides and, instead of identifying each labeled nucleotide one at a time as fragments pass through a gel, use a camera or other device to chronicle which bases are added in real time for hundreds of thousands to millions of fragments simultaneously.

Once the fragments are sequenced, the resulting reads have to be assembled correctly in order to reconstitute the original genomic sequence. This is done by sorting through all the reads to find sections that can overlap. Once they are all lined up, the consensus sequence can be created. In the HGP, the chromosome maps helped with this process because researchers knew the physical location of a fragment before they sequenced it. But once shotgun sequencing became the dominant methodology, scientists began to rely completely on computers, using specific biological algorithms called bioinformatics (now a growing field in itself), to do it all.

"The birth of genome bioinformatics was because of both higher throughput and shotgun sequencing. We used to have to assemble hundreds of reads and now it's millions or billions," says Patrick Chain, a bioinformaticist at Los Alamos. Most next-generation sequencers use short reads, increasing throughput but requiring more bioinformatics to assemble. One notable exception uses single fragments (no amplification) and produces very long reads but, lacking redundancy, is more prone to errors.

"Reads," or sections of DNA produced by sequencing, are lined up according to their overlapping sections to create a consensus sequence of the entire genome.

Los Alamos genome scientists have become experts at creating complete, high-quality genome assemblies by combining the results of multiple sequencing platforms. Lab scientists also analyze these sequences and investigate the function of their genes for a diverse set of research problems, such as identifying the culprit in a disease outbreak or understanding the subtle differences between species.

## Three billion's company

After the human genome was published, the NIH took on the challenge of deciphering the human data into more meaningful information, and the DOE, including its national laboratories like Los Alamos, dropped out of human genetics altogether. The DOE had created the Joint Genome Institute (JGI) in 1997 to unite the sequencing capabilities at DOE labs (Los Alamos, Lawrence Berkeley, and Lawrence Livermore), and in 2000 the JGI began to focus on sequencing microbes as related to DOE mission areas in carbon cycling, clean-energy generation, and environmental characterization and cleanup. This translated into years of work sequencing hundreds of microbes, which, when added to the public GenBank database, contributed to numerous comparative analyses.

One of the reasons microorganisms are interesting and valuable to study is that they live in a wide range of environments and are usually found in large, interdependent communities—interdependent upon each other and their surroundings. A huge variety of microbes in the soil, for example, can work in concert to degrade organic waste. Microbes are also important partners for other living organisms because of their ability to carry out functions to benefit their host environment, as is the case for the microbial community within the human gut that enables healthy digestion. In fact, microbes do quite a lot for humans, and there are about 10 times more bacterial cells on and in the human body than actual human cells.

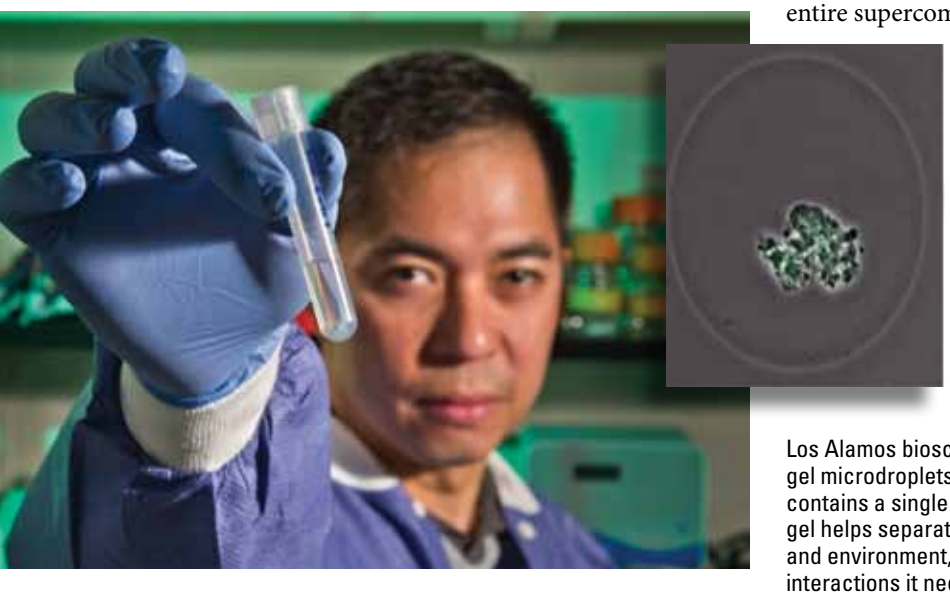
Although this interdependency makes them interesting, it also makes them very difficult to study because many rely on one another and their environments in order to grow. For that reason, isolating any specific type of microbe in a lab for sequencing can be difficult or impossible. The approach used to study these complex communities of microbes is called metagenomics, named for the idea of grabbing all the DNA in the community and treating is as one large *meta*genome. Metagenomics has become much more practical with modern sequencing.

"Now we can query more complex systems," says David Bruce, a manager for the genome group at Los Alamos. "Sequencing is no longer the rate-limiting step; now it's analysis."

In metagenomics, the challenge is to reassemble the billions of pieces that come from a wide variety of different organisms. Here, if some pieces reveal themselves to be from a known organism, scientists can assemble that organism's genome, but another tactic is to screen the pooled data for clues about what types of organisms are present. This is often achieved by organizing genes by certain known functions, rather than by organism. In addition to their arsenal of assemblers and other bioinformatics tools, Los Alamos scientists have designed two unique programs to tackle this challenge of classifying metagenomic data: Sequedex and GOTTCHA (Genomic Origins Through Taxonomic CHAllenge).

Both programs operate on the premise that specific kinds of signatures can help narrow the search criteria when comparing to known sequence databases. Sequedex focuses on amino acid signatures that are conserved—that is, common to two or more organisms or species. GOTTCHA focuses on nucleic acid (DNA or RNA) signatures that are unique to a genome or to a taxonomic group and tosses out all redundant genomic data. Both programs greatly reduce the amount of data that needs to be searched for matches, thus speeding analysis to the point that it can be carried out on a powerful laptop instead of an entire supercomputer.

Most sequencing requires a fair amount of DNA, so microbes have to be cultured and coaxed to replicate into a colony containing billions of cells. But since not all microbes can grow by traditional cultivation methods, researchers have developed new, potentially culture-independent techniques to isolate single cells from a mixed culture for sequencing. This approach, called single-cell genomics, is also promising for metagenomic samples—when a researcher may want



Los Alamos bioscientist Armand Dichosa holds up a vial of gel microdroplets. (Inset) Each tiny microdroplet (large circle) contains a single microorganism (green-gray shape). The gel helps separate the microorganism from its community and environment, while not completely isolating it from the interactions it needs to survive and grow.

# Better Living Through Biology

CREDIT: ROCKY MOUNTAIN LABORATORIES, NIAID, NIH

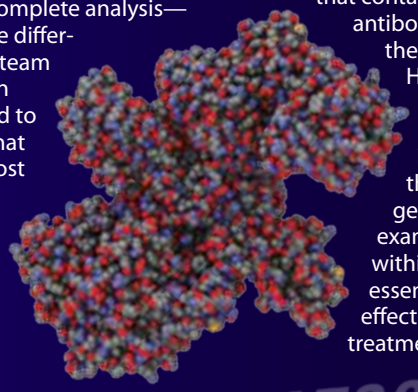
### *Identifying the cause of an outbreak*

During the 2011 *E. coli* outbreak in Germany, Los Alamos scientists provided data analysis to prove the origin of the bacterial strain and why it was so virulent. The strain of *E. coli* carried a toxin normally found in Shigella bacteria and looked similar to one from a 2009 *E. coli* outbreak in the Republic of Georgia. By doing a complete analysis—comparing base-by-base differences—the Los Alamos team showed that the German strain was indeed related to the Georgian one, but that the German strain had lost the genes for one type of Shigella toxin and picked the genes for another.

### *Bad botox*

Los Alamos scientists have characterized an extensive collection of the spore-forming bacteria *Clostridium botulinum*. This information assists public health officials in determining the specific type of botulinum neurotoxin that causes botulism in a patient.

"About 100 cases of infant botulism occur within the U.S. each year, and infants recover when provided a pharmaceutical product that contains a mixture of antibodies that neutralize the toxin," explains Karen Hill, the Los Alamos biologist who leads the project. "Identifying endemic strains that recur in certain geographic areas and examining the variation within the toxin types is essential for providing effective therapeutic treatment."

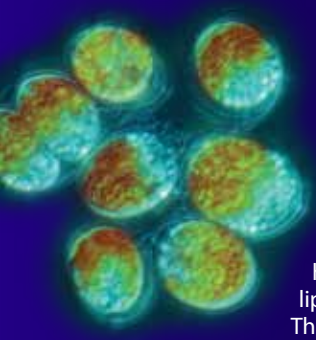


### *Soil secrets*

Using metagenomics, Los Alamos scientists have led an extensive effort to study microorganism populations in various types of soil—from forests and arid lands to arctic permafrost. Many soil microbes are responsible for the degradation of organic matter (dead plants and animals), which essentially takes carbon out of the ground and puts it back in the atmosphere, a major part of the carbon cycle. The research team is working to determine the impact of climate change factors on these soil carbon cycling communities in the context of other ecosystem factors, such as soil and plant type, or regional climate.

DNA Genomics

ATCGGATCTAGGGATCCCATGCTAGCTAGCTCGATCGA
TAGCCTAGATCCCTAGGGTACGATCGATCGAGCTAGCTACGC



### *Biofuels production*

Los Alamos microbiologists have been taking advantage of transcriptomics (sequenced RNA, indicating DNA that is currently in use) for their research towards creating biofuels from algae. By examining RNA transcripts from various strains of algae, at various time points, and under various conditions, they hope to identify genes involved in biomass or lipid production. (Lipids are used to make fuel.) They can then attempt to enhance growth and lipid production by adding these genes in the parts of the genome they know will be transcribed more often than others.

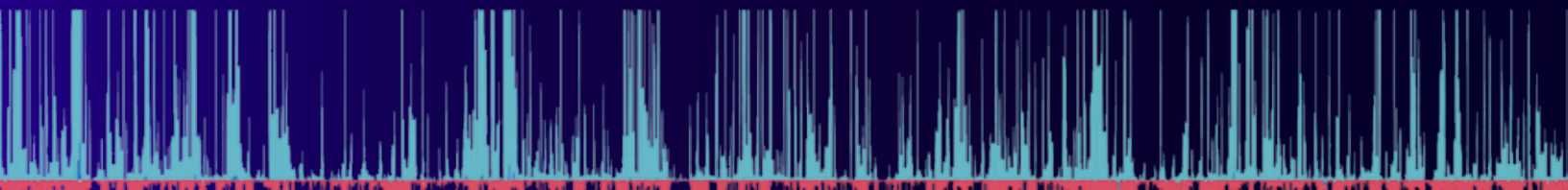
### *Personalized medicine*

The Los Alamos genome team is collaborating with the University of California, Davis, to understand the molecular mechanisms that cause cancer cells to become resistant to chemotherapy drugs. Cancer is defined as the uncontrolled growth of abnormal cells, which is often caused by genes that are not being regulated properly. Since most regulatory molecules are coded by RNA, sequencing and quantifying the RNA of a tumor can help researchers understand what has gone awry in those particular cells. This process also narrows down the list of genes that might be responsible for chemotherapy resistance—expediting further experimentation.

RNA Transcriptomics

### *Examining infection*

Transcriptomics can be used to examine interactions between organisms, such as when an organism has been invaded by a pathogen. The infection changes which genes are turned on or off by both participating organisms, and learning about these regulatory mechanisms can help scientists identify new targets for drugs or vaccines. To this end, Los Alamos genome scientists are examining the transcriptome of the bacterium *Yersinia pestis* (which causes plague) while it tries to evade destruction by a macrophage (an immune-system cell that engulfs invaders). The graphic below shows expression data from *Y. pestis* during this interaction; the red and blue spikes show the quantity of specific genes being expressed, giving an inside peek at the pathogen's defense mechanism.

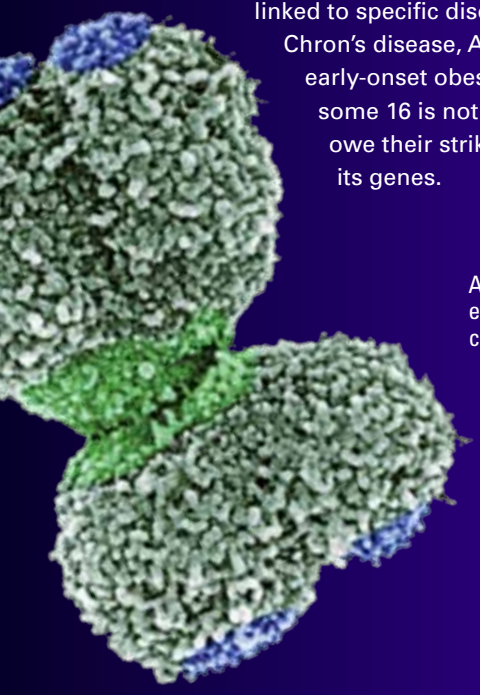


# Chromosome 16,
## *Where Are You Now?*

A major responsibility of the Laboratory during the HGP was to sequence the 90.4 million base pairs of chromosome 16 and to construct a detailed, high-resolution map showing the location of each of its approximately 1300 genes. The task was especially challenging because the chromosome has a larger-than-average proportion of duplications—long sections of DNA that are repeated on other parts of that chromosome or others.

Los Alamos biologist Norman Doggett, who was one of the principal investigators mapping chromosome 16, published a paper in 2006 (after the HGP had ended) about one of its duplications in particular, encompassing the HYDIN gene, which his team found to also exist on chromosome 1 (where it was named HYDIN2). It is one of the largest duplications across chromosomes in humans. Since the discovery of the HYDIN duplication, the HYDIN2 gene has been found to be associated with abnormal growth of the brain, causing developmental and behavioral problems, and the original HYDIN gene has been found to be responsible for primary ciliary dyskinesia, a rare genetic disorder that prevents the lining of the repiratory tract from removing mucus, exacerbating respiratory infections and conditions.

During the course of mapping chromosome 16, Los Alamos scientists contributed to the identification of several other important genes on the chromosome, including genes responsible for autosomal dominant polycystic kidney disease, Batten disease, and familial Mediterranean fever, to name a few. Today, more than 180 genes on chromosome 16 have been linked to specific disorders such as Chron's disease, Autism, and severe early-onset obesity. But chromosome 16 is not all bad; redheads owe their striking locks to one of its genes.



A colored scanning electron micrograph of chromosome 16.

to assemble an individual organism's entire genome even though the organism can't be cultured in isolation.

However, reliably assembling the complete genome from a single cell remains elusive; more DNA is needed. To this end, new Los Alamos technology allows the organism to be partially isolated in a droplet of gel for genomic study. The gel is porous enough for cell-signaling molecules (analogous to humans hormones) to flow in and out, such that the microbe is still able to communicate with other microbes and obtain nutrients from its environment. The captured cell grows into a microcolony of up to hundreds of identical cells while the gel keeps the microcolony sufficiently packaged for researchers to isolate the cells and assemble complete (or nearly complete) genomes in a high-throughput fashion.

### DNA in action

Although DNA encodes the instructions for an organism's full potential, not all of the DNA is transcribed, or copied into RNA, so that it can be used. Liver enzymes, for example, are encoded by DNA in all human cells but are only transcribed in the liver. Therefore, the transcribed RNA tells researchers about the part of the genome that a cell is currently using, known as its transcriptome. The RNA is assembled with DNA as a template and contains all the information needed for subsequent translation into a protein.

By sequencing the RNA transcripts, researchers can find out more information about what a cell is doing and when, instead of just what it has instructions to do. Teasing out the tiny RNA molecules from a mass of genomic data has always been tricky, but the high throughput of some next-generation technologies, coupled with new bioinformatics and statistics, is making it easier. Current transcriptomics projects at Los Alamos include analyzing algae strains for enhanced production of biofuels, studying the molecular mechanisms of how cancer cells become resistant to chemotherapy drugs, and understanding the complex interactions between pathogens and their host organisms.

It's difficult to overstate the opportunities to advance biological research enabled by the explosion of technology since the official end of the HGP. Just ask Los Alamos genome scientist Momchilo Vuyisich, who has been studying the transcriptome from *Yersinia pestis* (the bacterium that causes plague) as it tries to outwit an immune-system cell. He explains that being able to examine the changes in gene expression during this interaction is a huge breakthrough. "If you showed this *Yersinia pestis* data to someone 10 years ago," he says, "they would think you were in a fantasy world."

LDRD

*—Rebecca E. McDonald*